LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TH-698797

Analysis of Pre-Trained Deep Neural Networks for Large-Vocabulary Automatic Speech Recognition

G. A. Stevenson

July 28, 2016

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Analysis of Pre-Trained Deep Neural Networks for Large-Vocabulary Automatic Speech Recognition

Garrett Stevenson - ISCR, SULI, COMP CSS

Abstract

The structure of automatic speech recognition (ASR) applications has recently transitioned from using Gaussian mixture models (GMMs) to multi-layer, deep neural networks (DNNs) for acoustic modeling. Given the acoustic input, a DNN outputs probabilities of hidden Markov model (HMM) senone states. DNNs are initially pre-trained, then fine tuned and trained on appropriate data extensively to model high order features of speech input. This paper provides a pragmatic approach to identifying deficiencies and proposing areas of possible improvements in regard to context-dependent DNN-HMM architectures applied to large-vocabulary continuous speech recognition (LVCSR) tasks. DNNs are heavily reliant upon input data that is representative of the most important features of speech without redundancy. But the traditionally used signal processing method for speech has recently been shown computationally inefficient and less effective than more primitive methods. No sophisticated solution currently exists for speech signal feature detection, which would lead to improved data representation and thus performance. During DNN training, the fine tuning process is highly impactful on performance, but the exploration of variations to DNN fine tuning for speech has been limited. An optimal fine tuning algorithm for LVCSR has not been widely accepted, indicating there remains performance improvement to be found. As the breadth of training data used in a single DNN instance increases, the recognition error rate decreases, thereby providing evidence there is a pattern across all language ASR systems seek to model. The importance of understanding this general pattern of language is especially important to large vocabulary tasks. The DNN-HMM architecture has several components that are not yet optimized, indicating recognition error rates can be further improved upon.

Index Terms: Artificial neural network-hidden Markov model (ANN-HMM), context-dependent, acoustic modeling, large-vocabulary speech recognition (LVSR), pre-training, fine tuning.

The task of automatic speech recognition is heavily reliant upon efficient machine learning algorithms. Within the last five years, a major development in ASR was the finding that a deep neural network architecture could significantly outperform the traditionally used GMM in predicting emission probabilities of hidden Markov model states (HMM) representing phonemes. The superiority of DNN-HMM systems to GMM-HMM systems for ASR has been shown on a number of speech recognition tasks, ranging from well-known benchmarks to challenging large-vocabulary continuous speech recognition. [1, 2, 4, 6] The advantage of the DNN comes largely from its feature vectors [3], which are incrementally developed across consecutive speech frames. There are many different methods to pre-train, train, and optimize DNNs for ASR applications based on the format and density of training data, hardware limitations, and network architectures. In this paper, the speech recognition process from data representation to phone prediction is analyzed with a focus on deficiencies and possible areas of improvement. The studied DNN scope is limited to the current industry state of the art structure and therefore specific to context-dependent pre-trained deep neural networks, especially as applied to LVCSR. [3, 4]

A perceptron is an algorithm that classifies a vector input by its weighted summation. Deep neural networks are (L + 1) multi-layer perceptron (MLP) designed to model the posterior probability $P_{s|o}(s \mid o)$ of a HMM state s given the input observation vector o. [3] Each layer $l_i = 0, \ldots, L - 1$, is a hidden layer using the logistic function or hyperbolic tangent [1] to weigh and classify its input from the hidden layer above l_{i-1} which is then utilized by the layer l_i below. A hidden node z_j has an activation f(x)

$$(1)\ z_j = f\left(b_j + \sum_{k=1}^{p} w_{jk} y_k\right)$$

where w := $\{w_{jk}\}$ are the input connection weights, b_j is the bias of a unit j, and k is the index of input units 1,...,p corresponding to all the connections from the previous layer. [5] Using the logistic function, a hidden node's output to the next layer can be characterized as

$$(2)\ \ y_j = \mathit{logistic}(z_j) = \frac{1}{1 + e^{-z_j}}$$

where j is an index of all nodes in a layer and where the final layer L is the DNN's output (usually a softmax layer).

Before speech recognition can begin, input audio must first be processed. While this field is not part of the DNN architecture itself, signal processing methods have recently been explored to show significant differences in the final error rate of DNN ASR systems. Traditionally, DNNs rely upon Mel frequency spectral coefficients (MFCCs) as their input representation. The MFCC feature extraction approach provides good discrimination and a small correlation between components. [10] However some of the drawbacks to using MFCC are that it: performs poorly in noisy environments, is somewhat more computationally expensive compared to other methods, and is limited in its representation of signals by the spectrums it employs. New experiments at Microsoft have shown that reverting to more primitive methods of speech signal processing can actually be beneficial in DNN implementations receiving their input from the same source. [6] This is extremely significant, and while MFCC is still the default processing method because of its handling of microphone gain, this has opened the conversation to what known or unknown method might further benefit DNN performance. While MFCCs are not a limiting factor in ASR, tailoring the speech signal processing to provide DNN's with a maximum amount of relevant information and a minimal amount of redundant information has definitely been uncovered as a possible area of improvement in ASR.

DNNs are dependent upon learning from training examples and in fact, common practice is to iteratively pre-train and fine tune each layer of a DNN by initializing connection weights prior to processing training data. Individual layers are pre-trained with a semi-supervised learning algorithm(currently restricted Boltzmann machines), where they seek to learn nonlinear transformations that emphasize the main variations of their input data. [8] The importance of this pre-training has been verified [7, 9] and shown beneficial even in light of a large set of labeled examples. Semi-supervised pre-training acts as a regularizer, and therefore pushes a DNN model towards basins of attraction of minima [8] in order to support more abstract generalization of the training data set, because stochastic gradient descent and the top few layers of a DNN are prone to overfitting, even given significantly large amounts of training data. Although pre-training and fine tuning only make up about 10% of the training process, recent studies show the fine tuning process to be significantly more impactful on DNN error rate than pre-training, which brings up a few limiting features of DNN architecture. DNN's with a high number of hidden layers and nodes are difficult to optimize. Pre-training a layer l_i requires output from the input layer l_{i-1} therefore handicapping the ability to parallelize the process. In addition to this, DNN's with full connection from one layer's nodes to the next are initialized with weights that are random across a small range in order to vary the gradient input and thus diversify the nodes across a layer.

These two aspects of pre-training are known to need improvement, but their overall impact on DNN-HMM performance is generally assumed to be less than or near 0.5%.

Fine tuning a DNN follows pre-training and has a significantly larger impact on error rate reduction. [7,8,9] This method consists of iteratively adjusting the weights of connections in a DNN, minimizing the difference between the actual output vector and desired vector. Further detail on the fine tuning process can be found as it was originally presented in [11]. This backpropagation algorithm results in a DNN's hidden units representing dominant features of the speech domain in order to identify consistencies. The benefits of fine tuning are so great that some research currently advises doubling fine-tuning (labeled) data processing with respect to pre-training. [7] This process also represents a possible area of improvement in ASR. There currently exists no known optimal solution in finding the right sequence classification criterion for fine tuning. [1] Specifically with regard to large-vocabulary, context-dependent ASR, there currently is no proven optimal solution. This is partly due to the variety that currently exists in the deep architecture industry, and is certainly a field that should be further explored.

DNN training is the most explored area of ASR especially with respect to LVCSR tasks. A large factor that affects learning rates and hours spent training is the availability of training data well-fit for the desired ASR task. Data collection is costly, especially if the scope of desired application is more narrow than wide. Training data scarcity is especially limited when DNN-HMM systems are designed for languages less commonly spoken around the world. This may not appear to be a tragedy for the entire field of ASR, but [12, 15] show DNNs trained multilingually within a family of languages, post noticeably lower error rates than monolingual systems. This is confirmed in [6] but noted to increase training time even with the use of high performance clusters. [12] However, other studies in the area of training DNNs expand upon multilingual training by revealing a more general positive correlation between performance and any diversity of training data. Research within the last year has shown that the random introduction of certain types of noise into training sets, results in weaker tendencies towards overfitting and increased recognition in “noisy” environments. [13] Along the same lines, DNNs show improved performance after being trained with similar data from different sources [6], which can be explained by the same principle demonstrated in [13]. Because the diversification of training data has a positive effect on output but is limited (as all DNN training is) by overfitting, this aspect of ASR calls for further review, even in light of the parallel DistBelief system recently developed by Google [14]. Looking at DNN-HMM architectures as simply a pattern recognition system, provides insight into why software improves every time research moves in the direction of better understanding language. Multilingual and multi-source training have especially contributed to showing that there is an inherent pattern that exists in all language and therefore speech signals. As more information is found about the pattern that exists across all language, training ASR systems on the right combination of data will certainly give way to increased performance. What is unique about this field, is that it does not rely upon increased computing power or parallelism, but instead on a better understanding of human vocal communication.

A novel algorithm called “dropout” was proposed in [16] and recent experiments have supported its viability in helping prevent DNN overfitting. Dropout randomly omits a certain number of the feature detectors on training case. Dropout would replace (2) in a single instance with

$$y_j = f\left(\frac{1}{1-r} y_{j-1} * mw + b\right)$$

where *w* and *b* represent the weights and biases for the respective layer and m is a binary mask indicating which activation functions are not dropped out in the current case [17]. This technique helps improve generalization and the ongoing battle between a DNN and overfitting, and has been shown highly effective on a number of other non-speech tasks. [17] Yet in the context of LVCSR, a variety of training methods already exist (multilingual, noise, side-channel) that are known to improve DNN performance but make training a significantly longer task. Similarly, dropout increases training time by about a factor of two [17] and with sufficient LVCSR training already lasting around 4 days, dropout is therefore only useful in environments willing to sacrifice that time for reduced error rates.

A trained DNN's output is not the final result of an ASR system, but it instead supplies a HMM with the best acoustic modelling information it can provide to predict the target HMM states. HMMs have been popular models in LVCSR for a long time due to their flexibility, versatility, and consistent statistical framework. [18] A large contributor to the advantage of DNNs over GMMs for ASR is their ability to predict many thousands of tied triphone HMM states. [1] This creates a large increase in the number of HMM classes, but also inherently adds to the amount of training data and therefore time needed to initialize a DNN-HMM system. However, this is outweighed by the advantages of more information being supplied per window and the ability to use a triphone HMM decoder, which in a context-dependent environment, greatly reduces errors produced by phones that sound identical. This feature has been highly desired in the field of ASR and was originally thought to only be possible in speaker-dependent ASR application. But the speaker-independent DNN-HMM model outperforms its speaker-dependent versions by a large margin, indicating that the hidden layers of the DNN are progressively eliminating differences between speakers when trained diversely. [1, 2, 19] For this reason, it is no longer common to see new, adequately trained ASR applications ask users to repeat certain words or phrases before use, which is beneficial for the presentation of DNN-HMM software as complete and professional. While there could be improved performance benefits behind using increased power phone states (quad, quin, sex), the computational drawback would greatly outweigh any improved error rate.

The DNN-HMM system is a promising new development in the field of ASR technology. This architecture has improvements to be made in the three different levels of its structure, indicating DNN-HMM systems will be relevant for a significant period of time. At the signal processing level MFCCs have been outperformed by less sophisticated methods, opening up the field to finding a process optimal for speech signals in particular. Feeding a DNN-HMM system important, but non-redundant, data features is invaluable to the recognition process. The fine tuning process is a focal point of the DNN training process, but it remains unclear what sequence classification criterion is ideal for fine tuning an LVCSR focussed DNN. The underlying power in fine tuning DNN-HMM systems makes finding an optimal method a promising area for error rate reduction. Finally, diversification of training data in an LVCSR context has revealed speech patterns are not limited to their application language. This provides evidence that human language inherently has significant consistencies below the surface layer. As a deeper understanding of speech is developed, this low-level knowledge in particular could prove revolutionary to the field of ASR.

References:

[1] Hinton, Geoffrey, Li Deng, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Canhoucke, Patrick Nguyen, Tata N. Sainath, and Brian Kingsbury. "Deep Neural Networks for Acoustic Modeling in Speech Recognition." *IEEE Signal Processing Magazine* (2012): 82-97.

[2] Dahl, G. E., Dong Yu, Li Deng, and A. Acero. "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition." *IEEE Transactions on Audio, Speech, and Language Processing IEEE Trans. Audio Speech Lang. Process.* 20.1 (2012): 30-42.

[3] Pan, Jia, Cong Liu, Zhiguo Wang, Yu Hu, and Hui Jiang. "Investigation of Deep Neural Networks (DNN) for Large Vocabulary Continuous Speech Recognition: Why DNN Surpasses GMMS in Acoustic Modeling." *2012 8th International Symposium on Chinese Spoken Language Processing* (2012): n. Pag.

[4] Deng, Li, Geoffrey Hinton, and Brian Kingsbury. "New Types of Deep Neural Network Learning for Speech Recognition and Related Applications: An Overview." *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (2013): n. Pag.

[5] Titterington, D. M. "Bayesian Methods for Neural Networks and Related Models." *Statistical Science Statist. Sci.* 19.1 (2004): 128-39.

[6] Deng, Li, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, Yifan Gong, and Alex Acero. "Recent Advances in Deep Learning for Speech Research at Microsoft." *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (2013): n. pag.

[7] Yu, Dong et al. “Roles of Pre-training and Fine -tuning in Context-dependent Dbn -hmms for Real-world Speech Recognition.” (2010).

[8] Erhan, Dumitru et al. “Why Does Unsupervised Pre-training Help Deep Learning?.”*Journal of Machine Learning Research* 11 (2010): 625-660.

[9] Dubey, Avinava, Mrinmaya Sachan, and Jerzy Wiezorek. "Summary and Discussion Of: "Why Does Unsupervised Pre-training Help Deep Learning?"" *Statistics Journal Club* 36-825 (2014): n. pag.

[10] Anusuya, M. A., and S. K. Katti. "Front End Analysis of Speech Recognition: A Review." *Int J Speech Technol International Journal of Speech Technology*14.2 (2011): 99-145.

[11] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." *Cognitive modeling* 5.3 (1988): n pag.

[12] Heigold, G., V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean. "Multilingual Acoustic Models Using Distributed Deep Neural Networks." *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (2013): n. pag.

[13] Yin, Shi, Chao Liu, Zhiyong Zhang, Yiye Lin, Dong Wang, Javier Tejedor, Thomas Fang Zheng, and Yinguo Li. "Noisy Training for Deep Neural Networks in Speech Recognition." *EURASIP Journal on Audio, Speech, and Music Processing J AUDIO SPEECH MUSIC PROC.* 2015.1 (2015): n. pag.

[14] Dean, Jeffrey, et al. "Large scale distributed deep networks." *Advances in neural information processing systems*. 2012. n. pag.

[15] Ghoshal, Arnab, Pawel Swietojanski, and Steve Renals. "Multilingual Training of Deep Neural Networks." *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (2013): n. pag.

[16] Hinton, Geoffrey E., et al. "Improving neural networks by preventing co-adaptation of feature detectors." *arXiv preprint arXiv:1207.0580* (2012). n pag.

[17] Dahl, George E., Tara N. Sainath, and Geoffrey E. Hinton. "Improving Deep Neural Networks for LVCSR Using Rectified Linear Units and Dropout."*2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (2013): n. pag.

[18] Juang, Biing Hwang, and Laurence R. Rabiner. "Hidden Markov models for speech recognition." *Technometrics* 33.3 (1991): 251-272.

[19] Liao, Hank. "Speaker adaptation of context dependent deep neural networks." *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013.